图书情报工作
第65卷 第5期 2021年3月

利用迁移学习精准识别领域信息之探讨*

陆泉[1,2] 郝志同[1] 陈静[3] 陈仕[1] 朱安琪[1]
[1] 武汉大学信息资源研究中心 武汉 430072 [2] 国土资源部城市土地资源监测与仿真重点实验室 深圳 518034
[3] 华中师范大学信息管理学院 武汉 430079

摘 要：[目的/意义]将从互联网大数据中无监督学习的结果迁移到目标领域，解决目标领域因学习样本有限而信息识别效果难以提升的问题。[方法/过程]使用以中文维基百科等数据预训练的 RoBERTa 模型进行迁移学习，将学习结果映射到目标领域后使用 DPCNN 对其进行聚合凝练，然后结合部分标注数据微调模型完成领域信息的精准识别。[结果/结论]在10个领域内与未进行迁移学习的模型及经典模型 TextCNN 对比，提出的模型均较大幅度优于对比模型，平均后的精确率绝对提高4.15%、3.43%，召回率绝对提高4.55%、3.44%，F1 分数绝对提高4.52%、3.44%，表明利用网络大数据迁移学习可以显著提升目标领域的信息识别效果。

关键词：迁移学习 信息识别 RoBERTa
分类号：TP391.1
DOI：10.13266/j.issn.0252-3116.2021.05.011

1 引言

领域信息识别是自然语言处理（Natural Language Processing，NLP）中一个非常重要的研究方向，一直受到计算机科学、语言学等领域学者的关注，其目的是从不同领域文本集合中分离出特定领域相关的信息[1]。常规的手段是通过特征工程构造出一系列特征后使用机器学习进行处理，或者直接尝试训练 CNN（Convolutional Neural Network）、RNN（Recurrent Neural Network）等深度神经网络来学习用于信息识别的隐藏特征，这些方法的本质是从观测数据中构建规则并对观测外数据进行推理[2]，也就是说要想使模型推理得足够准确，就必须使用足够多的标记数据来学习模型。

为了保证模型性能，用于特定任务的大规模标记数据的收集成为一种刚性需求，这种需求对监督学习的许多应用场景提出了重大挑战，因为这些数据集往往需要人工标注，成本高、耗时长且容易出错。针对垂直领域的信息识别更是面临如此困难，领域自身数据少且标注困难成为制约模型性能发挥的关键因素。而与之相反的是，随着 Web2.0 的盛行，用户成为虚拟社区资料的重要创建者，每时每刻都在互联网上产生大量数据，其中不乏像维基百科这样的优质内容。

为了利用互联网大数据的优势去弥补领域数据不足的劣势，笔者使用 RoBERTa 预训练模型进行迁移学习，为了将迁移学习高度分散的输出结果针对领域信息进行凝练，笔者使用 DPCNN（Deep Pyramid Convolutional Neural Networks）完成对迁移结果的聚合与领域信息的判别。最后为了检验所构建模型的实际效果，在不同领域对模型进行实验验证。

2 相关研究

2.1 领域信息识别

领域信息识别是将所知的事物运动状态及其变化方式的形式或这种形式的某些特征参量与特定属性的"领域模板"的形式或它的特征参量进行比较，根据它们之间匹配情况的差别来判断该信息所应归属的领域类别[3]。广义上的领域信息识别包括文本识别、图像识别和语音识别，狭义上仅指针对文字信息的识别[4]，笔者所提的领域信息识别均指其狭义概念。

随着信息技术的飞速发展，各个领域的电子文本

* 本文系国家社会科学基金重点项目"心理账户理论视角下在线健康社区精准信息服务研究"（项目编号：20ATQ008）研究成果之一。
作者简介：陆泉（ORCID：0000-0002-8679-9866），教授，博士生导师；郝志同（ORCID：0000-0003-1803-2441），硕士研究生；陈静（ORCID：0000-0002-6444-2962），教授，博士生导师，通讯作者，E-mail：dancinglulu@sina.com；陈仕（ORCID：0000-0003-4664-7208），硕士研究生；朱安琪（ORCID：0000-0002-7526-1761），硕士研究生。
收稿日期：2020-07-03 修回日期：2020-09-19 本文起止页码：110-117 本文责任编辑：徐健

呈指数级增长且往往杂乱无序地分布于互联网社区内,对领域精准研究带来困难。仅靠人力难以去处理如此数量的信息,因此需要通过技术手段来实现领域信息的自动化识别。学者们在不断深入研究的过程中提出了大量经典的方法,这些方法可概括为两种类型:基于统计和机器学习的方法、基于深度学习的方法。

基于统计和机器学习的方法一般分为两个步骤,首先根据字词的统计量进行特征工程[5],接着利用机器学习算法在筛选出的特征上进行信息识别。如廖列法等[6]利用 LDA(Latent Dirichlet allocation)提取主题特征,采用 KNN(K-Nearest Neighbor)分类方法对稀土领域的专利文件进行识别,并达到了较好的识别精度。杨腾飞[7]等通过共词分析和 SVM(Support Vector Machine)在微博内对台风灾害信息进行了识别。基于统计和机器学习的方法曾在一段时期内推动了领域信息识别的飞速发展并取得了一系列优秀成果,但特征提取和识别算法分离的做法使得识别效果高度依赖于特征工程,特征的质量直接决定了模型的上限,这对人的能力提出了较高要求。所以随着神经网络的兴起,学者们逐渐将注意力转移到无需人工构造特征的深度学习模型上。

基于深度学习的方法对人脑的神经元进行了模拟,通常将目标和非目标领域的字词向量化表征后投入到层层神经网络中,利用损失函数调节神经元间的连接强度,实现对隐藏特征的区分,进而达成领域信息识别的目的。Y. Kim[8]较早将图像领域的 CNN 应用于文本信息,提出了 TextCNN 模型,该模型一经发布即刷新了 MR(Movie Review)等多个开源数据集的纪录,随后黄涛[9]利用该模型对不同领域的新闻信息进行了识别。之后又涌现出了 RCNN[10]、fastText[11]、DPCNN[12]等一批优秀模型,其中 DPCNN 更是在不增加太多计算开支的条件下便可捕获更长的文本依赖关系,在自然语言处理领域产生了深刻影响。目前,基于深度学习的方法是领域信息识别的主流方法,被广泛应用于学术界和产业界,但是深度学习方法存在严重的冷启动问题,需要大量的领域标注数据对模型参数进行调整。

2.2 迁移学习

迁移学习是一种利用已经相对成熟的领域(源领域)的知识来解决相关但未成熟领域(目标领域)问题的一种机器学习方法,它有效放宽了传统机器学习“学习过程需要大量带标注数据集,测试集与训练集需满足同分布假设”这两个前提[13]。迁移学习的出现为目标领域标注少甚至无标注问题的解决提供了方案。

迁移学习最早应用于计算机视觉领域,B. Zhou 等[14]在 ImageNet 和 Places 数据集上进行预训练,然后将迁移学习结果与小规模数据集结果进行比较,有力地证明了迁移学习可以取得更好的效果。

NLP 领域的迁移学习起步相对较晚,T. Mikolov 等[15]提出 word2vec,即利用大规模语料仅针对模型的第一层进行单词语义学习,该层可作为其他模型的词嵌入层直接使用,该方法产生了较大影响,但目标任务仍需从头开始训练。直到 2017 年 Google 提出 Transformer 架构[16],该结构在 NLP 领域具有里程碑式的意义,之后的迁移学习预训练模型几乎都基于 Transformer。2018 年,基于维基百科语料的预训练模型 GPT[17]和 BERT[18]出现,几乎刷新了所有的 NLP 任务榜单。之后学者们基于 BERT 又提出了性能更好的 RoBERTa[19]、ALBERT[20]等迁移学习模型,其中 RoBERTa 模型通过改进预训练任务、使用更大批次等方式取得了更优的效果,笔者使用该模型进行迁移学习。

2.3 迁移学习视角下的领域信息识别

从迁移学习的视角来看领域信息识别,主要有两个问题需要解决,首先是如何进行迁移学习,其次是迁移学习模型的输出结果如何使用。

对于第一个问题,随着预训练模型的出现学界和业界的研究者们逐渐达成共识,即使用预训练模型进行迁移学习,这样只要在预训练时投入足够丰富的语料,就可以完全避免从头开始训练用于迁移学习的模型,而这个时间动辄需消耗数月之久。目前,国外已经有少量学者开始使用预训练模型进行迁移学习从而完成领域信息的识别,如 N. Houlsby[21]等在 17 个公开数据集上对航空、经济、自然灾害等领域进行信息识别,发现使用预训练的 BERT 后识别性能至少提升 0.4%,T. Sharma 等[22]则基于 RoBERTa 预训练模型对食品信息获得了更好的识别效果。但是针对中文的迁移学习预训练模型出现较晚,所以还没有看到有国内学者使用其对领域信息进行识别。

对于第二个问题,现有的研究大多只是将预训练模型的输出通过 Softmax 层稍作修改后直接应用于具体的领域信息识别任务中,使用方式较简单却也能取得不错的效果。但是,用于迁移学习的预训练模型由海量数据通过无监督学习的方式训练出来,并不针对具体问题,其输出结果是一串高度分散的长序列,如果完全依赖预训练的输出而不进行更细化的调整,很容易对迁移学习到的结果造成浪费。

针对以上问题,笔者设计了“使用由大规模中文语料预训练出的 RoBERTa 模型进行迁移学习,通过 DPCNN 对迁移学习结果进行聚合凝练”的方案对领域信息进行识别,以期获得更佳的识别效果。总之,迁移学习还是一个新兴的研究领域,基于迁移学习的领域信息识别仍具有较大的提升空间,值得学者们展开进一步的深入研究。

3 基于迁移学习的领域信息识别模型构建

为了利用网络大数据的优势,笔者使用 2019 年刷新多项 NLP 任务记录的 RoBERTa 作为迁移学习的预训练模型。而迁移学习模型输出的结果是一个高度分散的长序列,为了捕获到输出序列的远距离依赖关系并对信息进行聚合,笔者使用 DPCNN 模型对迁移学习结果进行处理并对领域信息进行判断,最终设计的模型架构见图 1。

首先,将数据集按照字向量、分段向量、位置向量的三层结构进行嵌入式表征(Embedding),然后传递到

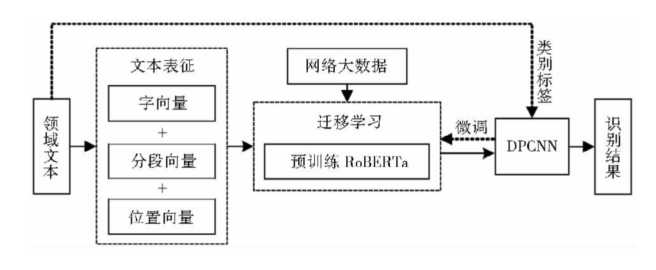


图 1 模型架构

使用中文维基百科等数据预训练的 RoBERTa 模型中,接着使用 DPCNN 对领域信息进行识别,并根据标注数据的反馈情况对预训练模型参数进行微调(Fine-Tuning),最终在包含众多领域信息的测试集内得到指定领域的信息识别结果,根据识别结果进而展开模型评估。

3.1 文本表征

模型的输入由字向量、分段向量、位置向量三层组合而成,如图 2 所示。第一层字向量是词表中每个 Token 的 Embedding;第二层分段向量用来区分输入文本的不同句子,同一句话的分段向量相同,如果输入单句则该层可全部置为 0;第三层位置向量记录每个字的相对位置和绝对位置。图中的[CLS]和[SEP]分别是输入文本的标记符和句子间的分隔符。

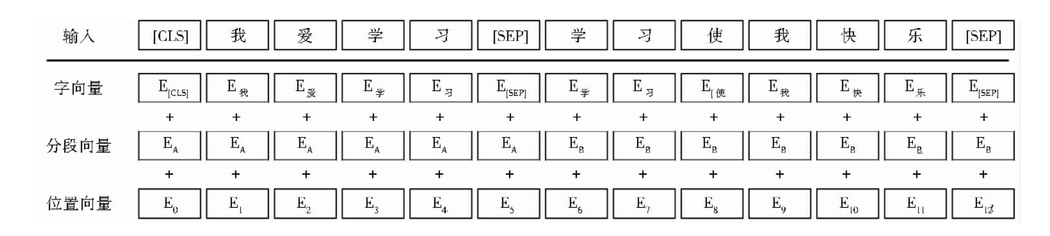


图 2 输入表征

上图中位置向量可按照公式(1)计算,其中 *pos* 为字词的绝对位置,*i* 表示 embedding 维度中的位置,*d* 表示向量维度。

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d})$$
$$PE_{(pos,2i+1)} = sin(pos/10000^{2i+1/d})$$
公式(1)

3.2 基于 RoBERTa 的迁移学习

RoBERTa 是 Facebook 在 2019 年发布的一个针对 NLP 任务的迁移学习模型,在 GLUE、SQuAD 和 RACE 3 个榜单上全部实现了最佳果。该模型在预训练时采用了 BERT 的遮蔽语言模型机制,但区别于 BERT 的静态遮蔽,RoBERTa 动态地对每次输入的序列进行随机遮蔽,然后基于上下文对遮蔽词进行预测,其遮蔽机制如图 3 所示。具体地,从每次输入序列中随机抽取 15% 做特殊处理,其中 80% 的概率被替换为[Mask],10% 的概率被随机替换掉,剩下 10% 的概率则保持不动。然后模型对 Mask 掉的字进行预测,虽然只对输入文本的一小部分进行了预测,但是在大规模语料的填充下,这并不会影响模型对语言的理解能力。

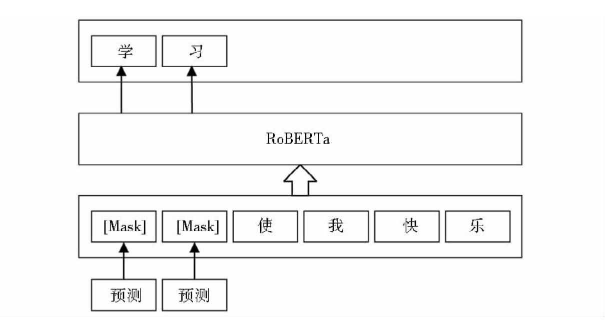


图 3 遮蔽语言模型作用机制

需要注意的是,这里使用的是全词遮蔽,即如果某个被 Mask 掉的字是构成词语的一部分,那么就将这个词的所有字均作遮蔽处理。这其实增强了模型处理复

杂问题的能力，就像上图中的例子，如果在知道“学”的条件下去预测“习”，那么无疑会比直接预测“学习”容易得多。

RoBERTa 模型的内部结构如图 4 所示，图中每个“Trm”都是一个 Transformer 的 Encoder 部分，从图中可以明显看到，在进行层与层间的递进时，任何一个“Trm”均使用注意力机制与上一层的所有“Trm”取得了联系，也就是说这种结构对信息是一种全方位的利用，而不像传统的 LSTM（Long Short-Term Memory）或 Bi-LSTM（Bi-directional LSTM）一样只能传递单向或双向的信息。

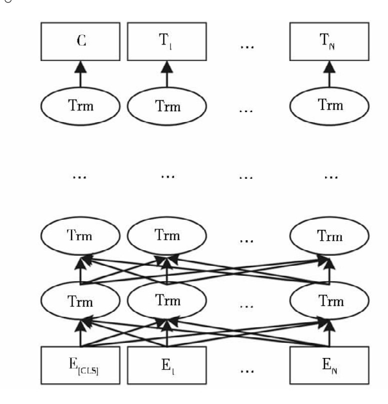


图 4　RoBERTa 模型结构

另外值得一提的是，图 4 的这种结构是一种高度并行化的结构，每个序列的节点生成同时进行，单个节点并不依赖于之前或之后的计算结果，所以基于该结构的 RoBERTa 模型可以快速地在相等时间内处理更多的语料信息，这也是它可以作为迁移学习模型发挥作用的关键原因。

3.3　基于 DPCNN 导向的微调与识别

迁移学习的结果是一个较长序列，并不能直接应用于领域信息的识别，必须根据任务的要求进行进一步的微调。为了捕获到该长序列中远距离节点的依赖关系，笔者使用腾讯 AI-Lab 于 2017 年提出的 DPCNN 对领域信息聚合后进行判断，再将判断结果反馈到迁移学习模型中进行参数微调，该过程见图 5。

DPCNN 将 RoBERTa 输出的 Embedding 连续投入到两个卷积层后进行 1/2 池化，然后对该过程进行重复，重复时为避免梯度消失对输入和输出使用残差连接。具体地，在每个卷积块后执行大小为 3 和步长为 2 的最大池化，这种池化策略将每个文档的内部表示的大小减少了一半，Feature Map 数量固定的情况下，每当执行 2 步下采样时，卷积核的有效覆盖率增加了一倍。因此，下采样周期过后，2 倍距离内的单词之间产

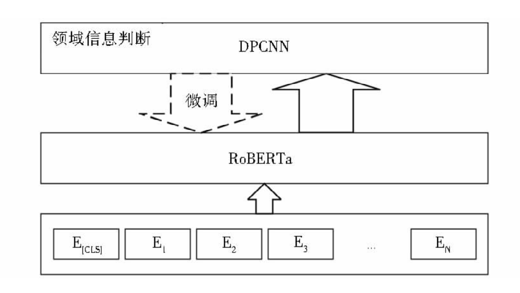


图 5　DPCNN 导向的微调示意

生关联，也就是说，DPCNN 以很高效的方式捕捉到了更远的信息，最终对全局信息进行了利用。

此外，用步长 2 进行下采样时，每个卷积层的计算时间减半（数据大小减半），从而形成一个“金字塔”，因此总的计算时间是由一个常数限定的，这个常数是最下层结构计算时间的两倍，这使得 DPCNN 在计算效率上也很有优势。

从图 6 可以更直观地看到，经过一系列深层的卷积和池化后，原始的输入序列被高度压缩，相当于对远距离信息进行了记录与聚合。图中每展开一次重复，上层节点探测到了更远距离的信息且序列长度缩短为原来一半，这样在对领域内容进行识别时可以对信息展开全方位的利用，从而提高信息的使用率。

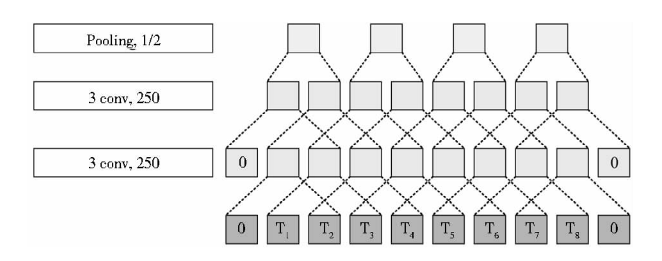


图 6　DPCNN 作用示意

最后根据领域信息的标注情况进行微调时，论文使用交叉熵损失函数和 Adam 优化器对整个网络的权重进行调节，交叉熵损失如公式（2）所示，然后在测试集上使用调整后的最终模型对领域信息进行识别并完成模型评估。

$$J(\theta) = -Y^T \log h_\theta(X) - (E-Y)^T \log(E-h_\theta(X))$$　公式（2）

上式中，Y、X 分别表示标签和处理后的变量矩阵，h_θ 代表 sigmoid 激活函数。

4　实验及结果分析

4.1　数据集

本次实验使用由胡文星二次整理的清华大学自然语言处理实验室发布的 THUCNews 数据集[23]，该数据

集采集自新浪网站，由财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐十大领域的文本数据构成，每个领域各包含 20 000 条数据，总计 200 000 条，实验按照“训练集：验证集：测试集 = 18：1：1”的比例对数据集进行划分后使用。用于迁移学习的 RoBERTa 预训练模型[24]由哈工大讯飞联合实验室发布，该模型使用了中文维基百科（https://dumps. wikimedia. org/zhwiki/latest/）和问答数据等（BQ corpus、CHNSENTICORP、CJRC、CMRC2018、LCQMC、MSRA、PFR、XNLI）通用语料进行训练。需要注意的是，维基百科由全球知识贡献者们编辑而成，数据质量可以得到有效保证，但互联网是一个相对开放的平台，其上产生的数据往往包含许多噪声信息，直接使用势必对模型精度造成影响，故模型在预训练过程中对所使用的补充语料进行了严格筛选，即上述补充语料均为 NLP 领域广泛使用的公开数据集，这些数据集由研究者们精心整理而成，具有较高的质量，这样便可以在很大程度上避免低质量数据带来的负面影响。

实验使用的硬件信息如下，GPU：Quadro RTX8000；显存：48G；CPU：2 × Xeon Platinum 8160；内存：128G。软件环境如下，系统：Ubuntu 16. 04 LTS，NVIDIA 驱动：418.56，CUDA 版本：10.1，编程语言：Python 3.7，深度学习框架：PyTorch1.5。

4.2 实验方案

为观察笔者所构建模型对领域信息识别的实际效果，同时设计了 3 组对照实验：①使用仅有训练集数据预训练的 RoBERTa 模型（即未进行迁移学习）并配合 DPCNN 对领域信息进行识别；②仅使用迁移学习后的 RoBERTa 对领域信息进行判断；③使用经典的深度学习模型 TextCNN。最终形成的实验框架如图 7 所示：

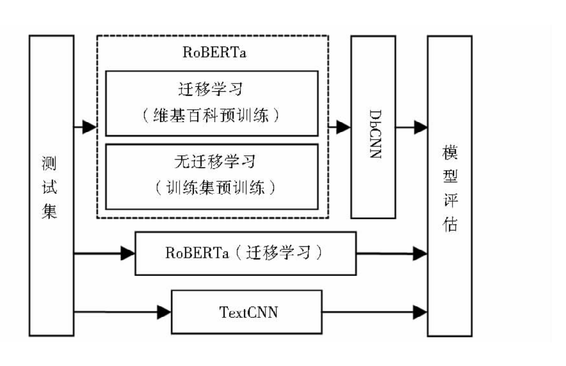


图 7 实验框架设计图

最后采用精确率、召回率以及 F1 分数作为评价指标对模型的信息识别结果进行评价。其中，精确率是指模型判断为正样本的数量中真实正样本所占的比例，召回率是指真实正样本中被模型正确识别出的比例，F1 分数是二者的调和平均数。3 个指标的计算公式如下：

$$精确率\ P = TP/(TP + FP) \qquad 公式(3)$$

$$召回率\ R = TP/(TP + FN) \qquad 公式(4)$$

$$F1 = 2P * R/(P + R) \qquad 公式(5)$$

公式（3）和（4）中，TP 表示将真实正样本预测为正样本的数量，FP 表示将真实负样本预测为正样本的数量，FN 表示将真实正样本预测为负样本的数量。

4.3 实验结果分析

在测试集上使用笔者所构建模型和其余基线模型分别对各领域信息进行识别并评估，将各模型的精确率、召回率和 F1 分数进行计算、统计后汇总成表 1。表中“TL”和“NoTL”分别表示经过迁移学习和未经过迁移学习的 RoBERTa 模型，“D”为 DPCNN 模型。

表 1 模型性能评估

领域	精确率				召回率				F1 分数			
	TL + D	NoTL + D	TL	TextCNN	TL + D	NoTL + D	TL	TextCNN	TL + D	NoTL + D	TL	TextCNN
财经	**0.932 1**	0.931 0	0.929 0	0.912 2	**0.934 0**	0.863 0	0.927 0	0.893 0	**0.933 1**	0.895 7	0.928 0	0.9025
房产	**0.962 9**	0.932 4	0.947 5	0.908 0	**0.959 0**	0.911 0	0.957 0	0.947 0	**0.960 9**	0.921 6	0.952 2	0.927 1
股票	**0.902 6**	0.895 7	0.889 8	0.875 8	**0.899 0**	0.790 0	0.896 0	0.839 0	**0.900 8**	0.839 5	0.892 9	0.857 0
教育	0.968 0	**0.970 8**	0.962 1	0.959 4	**0.969 0**	0.932 0	0.965 0	0.945 0	**0.968 5**	0.951 0	0.963 6	0.952 1
科技	**0.901 8**	0.761 7	0.870 9	0.864 8	**0.909 0**	0.898 0	0.904 0	0.870 0	**0.905 4**	0.824 2	0.887 1	0.867 4
社会	**0.935 1**	0.874 6	0.933 1	0.903 2	**0.951 0**	0.928 0	0.914 0	0.914 0	**0.943 0**	0.900 5	0.923 5	0.908 5
时政	**0.928 9**	0.861 4	0.917 2	0.880 2	**0.927 0**	0.926 0	0.908 0	0.911 0	**0.927 9**	0.892 5	0.912 6	0.895 3
体育	**0.990 0**	0.924 9	0.988 0	0.971 1	**0.989 0**	0.985 0	0.986 0	0.941 0	**0.989 5**	0.954 0	0.987 0	0.955 8
游戏	**0.976 0**	0.967 7	0.954 3	0.908 7	**0.935 0**	0.839 0	0.929 0	0.926 0	**0.955 1**	0.898 8	0.941 5	0.917 3
娱乐	**0.948 3**	0.910 6	0.945 4	0.919 5	**0.972 0**	0.917 0	0.970 0	0.914 0	**0.960 0**	0.913 8	0.957 6	0.916 8
均值	**0.944 6**	0.903 1	0.933 7	0.910 3	**0.944 4**	0.898 9	0.935 6	0.910 0	**0.944 4**	0.899 2	0.934 6	0.910 0

注：TL – 迁移学习；NoTL – 未迁移学习；D-DPCNN

在表 1 中,对比 < TL + D, NoTL + D, TextCNN > 这组实验,可以明显发现未进行迁移学习的 RoBERTa + DPCNN 模型实际效果要比经典的深度学习模型 TextCNN 稍差些,但使用迁移学习后各项指标都得到了显著提高。将表 1 中各领域内每组统计指标下的最高值进行加粗表示后可以明显看到,基于迁移学习的 RoBERTa + DPCNN 模型在各项指标下几乎都占据了榜首位置,对比经典的 TextCNN 模型,各领域的精确率、召回率、F1 分数平均后的绝对提高值分别为 3.43% 、3.44% 和 3.44% ;对比未迁移学习的 RoBERTa + DPCNN 模型,分别提高 4.15% 、4.55% 和 4.52% 。这些数据可以充分表明,在进行领域信息识别时引入迁移学习方法可以充分发挥大数据的优势,改善模型的识别性能。

另外也可以看到,对教育领域的信息进行识别时,仅使用训练集训练出的 RoBERTa + DPCNN 模型比迁移学习后的模型精确率还要稍高些,但继续观察召回率指标就可以发现,精确率的略微提高是以较大幅度牺牲召回率作为代价的,这对于领域信息识别的任务而言是较难接受的,因为会漏检相当一部分领域相关信息,所以综合下来其 F1 分数低于迁移学习后的模型。同时观察全部数据可以发现,"NoTL + D" 模型的召回率整体偏低,通过进一步的分析,笔者认为该模型是一个参数量很大的复杂模型,而训练集的样本数量较少,当所处理的问题同样比较复杂时整个模型的参数处于"欠学习"状态,所以不足以拟合到足够多的特征来对领域信息进行很好的判断。但是如果放到迁移学习环境中,因为有足够的语料来支撑模型训练,所以不存在这个问题。

继续观察 < TL + D, TL, TextCNN > 这组实验,可以进一步发现,相比于传统的深度学习模型 TextCNN,迁移学习后的模型在各项指标上几乎均实现了超越,仅使用预训练 RoBERTa 模型就在精确率、召回率、F1 分数 3 个指标上对 TextCNN 平均提升了 3.34% 、2.56% 、2.46% 。再仅观察 < TL + D, TL > 该组实验,发现使用 DPCNN 对迁移学习结果进行处理后,模型性能得到了进一步提高,精确率、召回率、F1 分数平均提高 1.09% 、0.88% 和 0.98% 。这说明对预训练模型的输出展开进一步的细化调整可以再次提升模型性能,避免对迁移学习的结果造成浪费。

此外,对上述的识别结果展开进一步的人工分析后有如下发现:①相比于未经过迁移学习的模型,迁移学习后的模型对文字中不包含明显领域特点的信息具有更好的识别效果。例如,时政领域下的一则新闻"我国将继续加大对传销犯罪的惩处力度",全文并未涉及政策、局势等相关字眼,但语义层面的确属于时政领域,最终只有 < TL + D, TL > 两组迁移学习模型识别出了该信息。②对于可能包含多个主题的领域信息,所有模型的识别结果均较差。例如"热门学科大揭秘:金融外贸类专业就业实情"这则新闻,新闻实际上是在分析热门学科,理应归属教育领域,只是该学科和金融有关,从而导致所有模型在对教育信息识别时均未识别出该则新闻,相反,在对财经信息进行识别时,所有模型均错误地将其识别出。

对领域信息识别而言,实际应用环境会更加复杂,目标领域所涵盖的信息数量要远低于真实空间内的信息总量,即二者的数据量处于高度不平衡状态。为了进一步评估模型的准确性和再不平衡样本空间下的泛化能力,笔者分别在 10 个领域内对各模型进行 ROC (Receiver Operating Characteristic) 曲线的绘制,见图 8。

图中模型后括号内的值代表 ROC 曲线覆盖下的 AUC(Area Under Curve) 面积,该值一方面可以作为模型准确性评估的参考,另一方面则在很大程度上代表了模型在非均衡数据集下的表现能力。从图中可以看到,测试的几个模型的 AUC 值均比较高,但总体上还是经过迁移学习后的模型在非均衡数据集上的泛化能力更好些。将每个模型在所有领域的 AUC 平均后,TL-RoBERTa + DPCNN、NoTL-RoBERTa + DPCNN、TL-RoBERTa 和 TextCNN 的 AUC 值分别为 0.990 、0.981 、0.985 和 0.986 ,同样表明笔者所构建的模型表现最佳,具有更好的泛化性能。

5 结语

笔者针对领域数据较难获取、在有限数据集上模型性能再难以提升的问题,提出一种利用迁移学习对领域信息进行更精准识别的方法。实验结果表明,使用迁移学习后的 RoBERTa + DPCNN 模型相比无迁移学习的模型和经典的 TextCNN 深度学习模型在性能上有了大幅提升,平均后的精确率绝对提高 4.15% 、3.43% ,召回率绝对提高 4.55% 、3.44% ,F1 分数绝对提高 4.52% 、3.44% ,充分证明了迁移学习的有效性。另外,使用 DPCNN 对迁移学习模型的输出进行聚合凝练后,精确率、召回率、F1 分数分别平均提高 1.09% 、0.88% 、0.98% ,说明对迁移学习结果进行更细化的调整有利于提高模型的识别性能。

论文也有不足之处,在对迁移学习模型进行微调

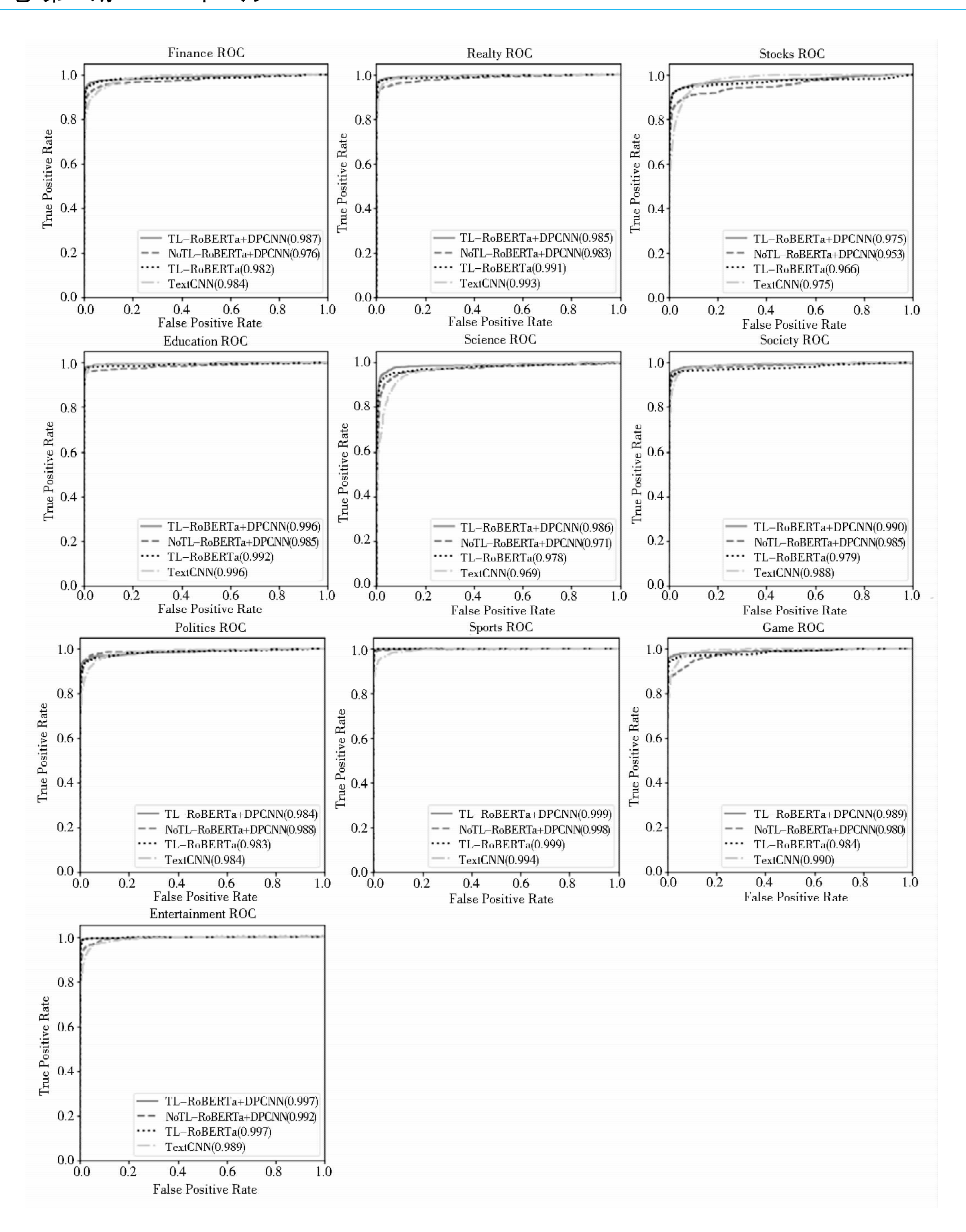


图 8 模型 ROC 曲线图

时，笔者保留了原预训练模型的全部参数，但是有研究[25]表明，针对性地对迁移模型的参数进行取舍可以在实际任务中获得更好的表现。下一步工作中，将针对此问题做更具体地研究，以期获得更佳的领域信息识别效果。

参考文献：

[1] RINGEL D, RADINSKY K, MARKOVITCH S. Cross-cultural transfer learning for text classification[D]. Israel: Technion, 2019.

[2] YU S, SU J, LUO D. Improving BERT-based text classification with auxiliary sentence and domain knowledge[J]. IEEE access, 2019, 7: 176600－176612.

[3] 潘洪亮，王正德. 信息知识词典[M]. 北京：军事谊文出版社，2002.

[4] 张学工. 模式识别[M]. 3 版. 北京：清华大学出版社，2010.

[5] MA Y, TANG J, AGGARWAL C. Feature engineering for data streams[M]//Feature engineering for machine learning and data analytics. Boca Raton: CRC Press, 2018: 117－143.

[6] 廖列法，勒孚刚，朱亚兰. LDA 模型在专利文本分类中的应用[J]. 现代情报，2017，37(3)：35－39.

[7] 杨腾飞，解吉波，李振宇，等. 微博中蕴含台风灾害损失信息识别和分类方法[J]. 地球信息科学学报，2018，20(7)：906－917.

[8] KIM Y. Convolutional neural networks for sentence classification [C]// YUVAL M. Empirical methods in natural language process-

ing. Qatar: ACL, 2014: 1746－1751.
[9] 黄涛. 基于机器学习的新闻分类系统研究与实现[D]. 北京: 北京邮电大学,2019.
[10] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]//IJCAI. Proceedings of the twenty-fifth international joint conference on artificial intelligence. New York: AAAI Press, 2016: 2873－2879.
[11] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]//MIRELLA L. the 15th conference of the european chapter of the association for computational linguistics. Spain: EACL, 2017: 427－431.
[12] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//HINRICH S. Proceedings of the 55th annual meeting of the association for computational linguistics. Vancouver: ACL, 2017: 562－570.
[13] 庄福振, 罗平, 何清, 等. 迁移学习研究进展[J]. 软件学报, 2015, 26(1):26－39.
[14] ZHOU B, LAPEDRIZA A, XIAO J, et al. Learning deep features for scene recognition using places database[C]//ROMAN G. International conference on neural information processing systems. Cambridge: MIT press, 2014: 487－495.
[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//JUN Z. Advances in neural information processing systems. Harrahs: NIPS, 2013: 3111－3119.
[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//ISABELLE G. Advances in neural information processing systems. Long Beach: NIPS, 2017: 5998－6008.
[17] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J/OL]. [2020－06－28]. https://arxiv.org/pdf/1810.04805.
[19] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J/OL]. [2020－04－01]. https://arxiv.org/pdf/1907.11692.
[20] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J/OL]. [2020－04－01]. https://arxiv.org/pdf/1909.11942.
[21] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[J]. [2020－04－01]. https://arxiv.org/pdf/1902.00751.
[22] SHARMA T, UPADHYAY U, BAGLER G. Classification of cuisines from sequentially structured recipes[C]//2020 IEEE 36th international conference on data engineering workshops (ICDEW). Dallas: IEEE, 2020: 105－108.
[23] 孙茂松,李景阳,郭志芃,等. THUCTC:一个高效的中文文本分类工具包[EB/OL]. [2020－05－10]. http://thuctc.thunlp.org/.
[24] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[J/OL]. [2020－06－01]. https://arxiv.org/pdf/2004.13922.
[25] TENNEY I, DAS D, PAVLICK E. BERT rediscovers the classical NLP pipeline[J/OL]. [2020－01－01]. https://arxiv.org/pdf/1905.05950.

作者贡献说明：
陆泉:提出研究思路,设计论文框架;
郝志同:模型构建与实验,撰写论文;
陈静:设计研究方案,修订论文;
陈仕:数据预处理;
朱安琪:实验设计。

Discussion on Using Transfer Learning to Accurately Identify Domain Information

Lu Quan[1,2] Hao Zhitong[1] Chen Jing[3] Chen Shi[1] Zhu Anqi[1]
[1] Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518034
[3] School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] To solve the problem that the identification effect of the target domain information is difficult to improve because of not enough samples, we will transfer the results of unsupervised learning from big data to the feature space of the target domain. [**Method/process**] Used the RoBERTa model, which was pre-trained with Chinese Wikipedia and other data, for transfer learning. After mapping the learning results to the target domain, DPCNN was used to aggregate and condense it, and then fine-tuned the model with part of the labeled data to complete the accurate recognition of domain information. [**Result/conclusion**] Compared with the model without transfer learning and the classic model TextCNN in 10 fields, the model in this paper is much better than the comparison models. After average, the precision is increased by 4.15% and 3.43%, the recall is increased by 4.55% and 3.44%, and the F1 score is increased by 4.52% and 3.44%. It shows that knowledge transfer using big data can effectively improve the information recognition effect in the target field.

Keywords: transfer learning information recognition RoBERTa